DOCUMENT RESUME

ED 138 624 TM 006 247

AUTHOR Deck, Dennis
TITLE Validity of Word Deletion Items as a Measure of Reading Comprehension.
PUB DATE [Apr 77]
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Age Differences; *Cloze Procedure; Context Clues; Elementary Education; Elementary School Students; *Reading Comprehension; *Reading Tests; Statistical Analysis; Test Construction; Test Reliability; *Test Validity

ABSTRACT

The construct validity of measures of reading comprehension was discussed. A word deletion procedure was proposed consistent with a model of comprehension which emphasizes the importance of context. The procedure assures sensitivity to context by allowing only highly constrained words to be deleted and by increasing the information value of context. Third and sixth grade students answered word deletion items at three levels of contextual constraint. The results indicated that while sixth graders were more skilled than third graders, both groups could use contextual information to reduce uncertainty. The group differences were not observed for items which were not highly constrained. That highly constrained items were more difficult and yielded a more narrow range of responses than less constrained items provided evidence of the construct validity of the proposed measure. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED138624

Validity of Word Deletion Items As A Measure of Reading Comprehension

Dennis Deck

Pennsylvania Department of Education

Presented to the Annual Meeting of the

American Educational Research Association

New York, New York

April, 1977

TM006 247

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Validity of Word Deletion Items As A Measure of Reading Comprehension[1]

Dennis Deck

Reading comprehension tests have been used for many years with relatively little concern for construct validity. Recently, however, educators have begun to question these measures. For example, critics have suggested that there should be systematic procedures for constructing items (e.g., Anderson, 1972; Bormuth, 1970). Generally, comprehension questions do not have a clear relation to the original passage or can be answered without a full understanding of the text. The finding that students can answer many questions on standardized comprehension tests without reading the supporting passage (Pyrczak, 1974) suggests that guessing, inference, and prior knowledge are also being tested. Attention has also been given to aspects of the task which place demands on other abilities such as memory (Carroll, 1972). Although important, these methodological threats to validity stem from a more basic problem, the lack of a clear model of comprehension. Without an explicit definition of the construct and its relationship with other constructs, it is difficult to evaluate the adequacy of any measure of comprehension.

The traditional comprehension model for test construction is based on the various lists or taxonomies of component comprehension skills. These taxonomies, however, have rarely been validated. They imply distinctions or relationships which are difficult to justify upon a

---

[1]This study is based on a doctoral dissertation submitted to the Department of Educational Psychology, Pennsylvania State University, 1975. The author is indebted to Francis J. Di Vesta, Chairman, J. Jackson Barnette, Paul Games, and John Nesselroade for their assistance and constructive comments throughout the project.

closer analysis. While Davis (1968) has provided some evidence of several component comprehension skills, reanalyses of the same data did not confirm his findings (Spearitt, 1972; Thorndike, 1973). The problems with the taxonomies and the failure to psychometrically distinguish between these skills suggests that the model has not been satisfactory for producing valid comprehension measures.

Another model for test construction was proposed by Bormuth (1970) based upon linguistics. The theory of transformational grammar (Chomsky, 1965) was used to operationalize procedures for constructing comprehension items. Bormuth argued that since surface structure can be changed without changing the deep structure or meaning of a sentence, one can construct comprehension items by lifting a sentence from a passage and using transformations to generate questions. The advantage of this approach is that it is based upon a clearly defined construct where there is some experimental evidence of its validity. On the other hand, psycholinguists have begun to question the adequacy of syntactical deep structure as an approach to semantics (Perfetti, 1972).

A third attempt to provide a model for testing comprehension is represented by the classic debate over how comprehension differs from reasoning. Carroll (1972), for instance, has suggested that we can draw a useful distinction between the "apprehension of linguistic information" and "relating linguistic information to a broader context." Alternatively, Bransford & Johnson (1973) have provided evidence of the strong influence of context in the interpretation of sentences to show that inference is an integral part of comprehension. Apart from the lack of clear empirical evidence for such a distinction (Thorndike, 1973), it is unlikely that

many test users would accept the more narrow definition of comprehension. At least some inference would seem to be necessary.

An alternative model of reading comprehension has grown out of the psycholinguistic and information processing theories (e.g., Bransford & Johnson, 1973; Levin & Williams, 1970; Smith, 1971). Here reading is characterized as an active process of reconstructing meaning based upon one's general knowledge of the world. It is not primarily a visual process, but rather a "psycholinguistic guessing game" (Goodman, 1967) in which the reader uses the context of the preceeding material and both linguistic and nonlinguistic knowledge to form hypotheses about the meaning of the passage. Then the reader selectively searches the text for further information to confirm and disconfirm these expectations. A major difference between the beginning and the fluent reader is how effectively context is used to activate the reader's linguistic and nonlinguistic knowledge to aid comprehension.

In the model, the ability to comprehend can be defined as a reduction of uncertainty; that is, the reduction of the set of alternative meanings hypothesized by the reader (Smith, 1971). The reader's ability to use context allows him/her to partition the range of different meanings of a word or sentence so that little additional information must be sampled from the text to confirm expectations. While this conceptualization does not capture all that is commonly included under the label of comprehension, it seems to clarify the role of inference and to suggest new ways in which comprehension can be measured and evaluated.

A measure of comprehension by this definition must be sensitive to the way in which the reader reduces uncertainty by eliminating alternative

meanings based upon contextual information; it must be contextually dependent. Procedures to measure comprehension must be developed and evaluated with emphasis on the degree to which the responses to the items are due to the constraint of context and not due to other factors.

The deletion of words in sentences has been proposed as an appropriate procedure for measuring the reader's use of contextual information (Smith, 1971) since the reader can only supply the missing word if he/she uses semantic and syntactic cues provided by the preceeding portion of the passage. The most popular word deletion procedure has been the cloze procedure (Taylor, 1957) which involves selecting a passage and deleting every nth word, usually every fifth. The traditional cloze procedure, however, is rather crude (Carroll, 1972). Responses to the deleted words are largely dependent upon local redundancy, i.e., the grammatical and semantic context in close proximity to the deleted word, rather than the more remote context of the preceeding passage. In fact, generally less than five words of context are needed to answer these questions (MacGinitie, 1961). The model adopted here suggests that contextual constraint in word deletion measures must be more carefully controlled to insure that a deeper understanding of the context is required before the reader can supply the deleted word. The test constructor must be able to show that a good reader can eliminate all but a few alternatives when context is present but not when context is absent.

To meet this criterion, words for deletion must be carefully chosen to minimize the importance of cues within the sentence or clause. Since function words are largely constrained by context, deletions should be

restricted to more semantically laden words such as nouns, verbs and adjectives. No word should be deleted which can be guessed from associations with other words or semantic cues in the rest of the sentence. For instance,

He mowed his lush, green _____.

would not be an acceptable item; the answer can be supplied without remote context. There must be a number of possible answers when only the sentence or a clause is presented.

Words in context must be chosen carefully to minimize the constraint or information value of context. The choice of one word or another to convey a particular meaning is determined by the inferred set of alternatives among which the reader will have to distinguish (Olson, 1970). That is to say that one would call a ball a "sphere" in geometry class but then call it a "toy" when telling a child to pick up his/her room. The alternatives partitioned in each case were different. To convey more information, the choice of words in context must limit the possible interpretations of some aspect of the target sentence. Thus, one might use the word "break" rather than "damage" since it implies an object that is characteristically rigid and brittle.

The purpose of this study was to validate the proposed word deletion procedure as a measure of reading comprehension. Viewing comprehension as the reduction of uncertainty leads to certain predictions about the responses of good and poor readers to such a measure. If readers use contextual information, their responses should demonstrate less uncertainty when more constraining contexts are provided. Older, more fluent readers should be better able than beginning readers to eliminate alternative

meanings using available contextual information. However, when contextual information is unavailable, this difference should no longer be observed. Thus, depending upon the degree of contextual constraint, there should be age differences in the uncertainty of responses to the deleted words. The number of words to which an individual responds correctly is taken as an index of an individual's ability to use contextual information and should also reflect age differences in comprehension ability. When there is little or no contextual information available, the test score reflects something other than this ability.

METHOD

Subjects

The subjects were 126 third grade and 122 sixth grade students from two small town elementary schools. The data from six students were eliminated because either the student failed to finish the test in the allotted 45 minutes or the student left more than four items unanswered. Two students were randomly eliminated to maintain equal cell n's.

Materials

The comprehension measure used in the study was a set 36 items, each consisting of a target sentence having one word deleted and a context sentence. The deleted word was one of three of the grammatical cases identified by Fillmore (1968): Locative, Objective and Action. Twelve items were written for each case.

Items were constructed by first generating a list of about 200 words for each of the three cases and randomly selecting words from each list to form the base for the items. For each item, three degrees of

Contextual Constraint were then generated. First, the target sentence and a highly constraining context sentence were written. For example:

Jack heard a loud bark during his walk.
He saw a _____ behind the barn.

A second level was formed by replacing one adjective or noun in the context sentence with a less constraining paraphrase such as "animal" for "bat." The above example was changed to:

Jack heard a loud noise during his walk.
He saw a _____ behind the barn.

The word "noise" places fewer restrictions on the word that will fit in the blank than the word "bark." A third level of constraint was formed by eliminating the context sentence, so the item above was changed to:

He saw a _____ behind the barn.

Several extraneous factors which might affect item validity were controlled. The content of the items involve common, daily experiences which did not require any special knowledge that might provide sixth graders with an advantage in comprehending the items. Only words found frequently in third grade readers, 20 words per 800,000 by the Carroll count (Carroll, Davies, & Richman, 1971), were used in constructing the target and context sentences. The grammar and length of sentences were controlled by using a simple, grammatical form for the context and target sentences throughout the item construction.

Each item was typed on a 3" x 5" sheet of paper with a 1 inch space for the deleted word. The items for each test book were shuffled to produce a different randomized item order for each subject. Four sample items preceded the experimental items for practice purposes. Since subjects could not receive more than one form of the same item, each

booklet contained all of the items from only one level of Contextual Constraint. Thus, there were three versions of the test, each containing all 36 items at one of three levels of Contextual Constraint (Low, Medium, or High).

Internal consistency reliability estimates using Coefficient Alpha indicated that the High constraint measure was reliable, .83 at the Third Grade and .85 at the Sixth Grade. The measure correlated highly with the Gates-MacGinitie Reading Test, r = .74 for Third Grade (n = 39) and r = .77 for Sixth Grade (n = 38).

Procedure

In classrooms of about 20 students, the three versions of the test were distributed randomly, with randomization recycled at each successive group of three students. Ss were told that they were about to take a reading test that was different from most tests they had taken. E read the four sample items instructing the Ss to think of the word that made the most sense in the sentence. After each item, the children were asked to write down their responses in the blank space and then a variety of responses were elicited orally from the class. The Ss were told to leave no item unanswered and to ask for help if they had any difficulty reading a word, finding an answer, or spelling their answer.

During the test, E answered any questions concerning reading an unfamiliar or lightly printed word and spelling an answer. If a S could not find an answer, E read the item aloud and asked the S to write down the first word that came to mind. When a S finished, the booklet was collected and the time recorded. Ample time was given for the task.

Dependent Variables

First, a measure of uncertainty was used to determine the constraining affects of context on the group responses to an item. The measure of uncertainty was H, the Shannon-Wiener measure of information or uncertainty (Attneave, 1959). This measure has been used as a general procedure for quantifying uncertainty or entropy in a variety of verbal tasks. It takes into account both the diversity of responses and the frequency of the particular response since both reflect the elimination of alternatives.

The frequencies of unique responses to individual items were tabulated and used to calculate H. Responses which differed only in tense and number were scored as equivalent. A separate H value was obtained for the three levels of Contextual Constraint in both Grades on each of the 12 items for each Case.

Second, the number of responses which were correct in context were tabulated to determine the affects on individual's responses. A response word was considered correct if it made sense in context and thus reflected the constraint placed upon it by the context sentence. Spelling and grammatical errors (e.g. tense and number) were ignored in the scoring. Two judges scored the test booklets with 93 per cent agreement.

RESULTS

To demonstrate the construct validity of defining reading comprehension as the use of contextual information to eliminate alternatives, two analyses of the responses were made. First, the effects of Grade and Contextual Constraint on the uncertainty of responses to the items were analyzed using items as the random factor. Second, the affects of Grade and Constraint on the number of correct responses made by the individual were analyzed using the Ss as the random factor.

Item Uncertainty

Using the item H values based on group responses, the results of a 3 x 2 x 3 mixed analysis of variance[2] reveal significant main effects for Grade, $F(1,33)=55.5$, $p < .05$, and Contextual Constraint, $F(1,33)=56.7$, $p < .001$. The interpretation of these main effects was precluded by a significant interaction between them, $F(1,33)=9.2$, $p < .01$. The main effect of Case was not significant, $F(1,33)=1.4$, $p > .05$.

The Tukey Wholly Significant Difference (WSD) test was used to determine the locus of the interaction (Winer, 1962). As Figure 1 shows, uncertainty of the responses to the items decreased as Contextual Constraint increased at both Third and Sixth Grades, though there was no difference between the mean H values for the Low and Medium constraint items. There was significantly less uncertainty in the Sixth Grade responses than in the Third Grade responses on Medium and High, but not Low, constraint items. This interaction indicated that response uncertainty decreases with age or ability, but only as the level of constraint imposed by context increases.

Number of Words Correct in Context

In order to study the effect of Contextual Constraint on an individual's responses, the number of responses each student gave which were correct in context were analyzed by a 2 x 3 x 3 mixed analysis of variance[3].

---

[2]As anticipated, the AOV assumption of homogeneity of variance and correlations between repeated measures was violated due to the effect of levels of Contextual Constraint on item uncertainty. The Box-Greenhouse-Geisser index (Winer, 1962) was used to conservatively adjust the df on within-item F's. The heterogenous cell variances replaced the pooled mean square error in the WSD follow-up tests.

[3]As anticipated, the AOV assumptions of homogeneity of variance and of covariance matricies of independent groups were violated due to the ease of answering Low constraint items. The df were conservatively adjusted using the Box-Greenhouse-Geisser index. To test the simple effects for the Contextual Constraints by Case interaction, the appropriate mean square error replaced the pooled mean square error.

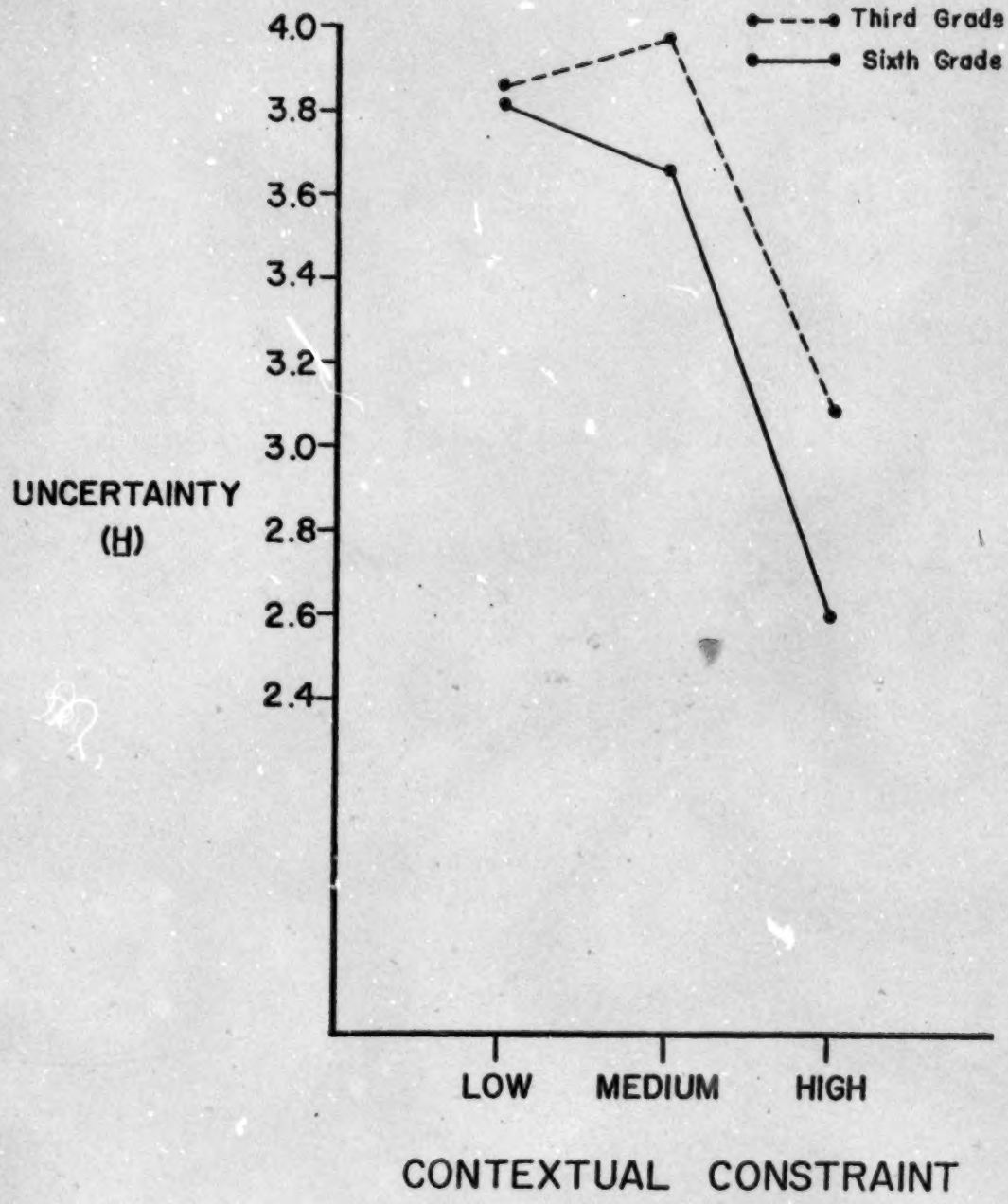

Figure 1. Mean item uncertainty by grades at three levels of contextual constraint.

The results indicate significant main effects for Grade, $F(1,234)=30$, $p < .01$, Contextual Constraint, $F(2,234)=94.6$, $p < .001$, and Case, $F(1,234)=9.6$, $p < .01$. Sixth Grade students performed significantly higher at all levels of Case and Contextual Constraint. The interaction between Case and Contextual Constraint while significant, $F(2,234)=15.5$, $p < .001$, will not be discussed since it accounted for only a small proportion of the total variance, $\omega^2=.02$ (Vaughan & Corballis, 1969). At each level of increasing Contextual Constraint, students made fewer correct answers. There were slight differences in the difficulty of items in the three cases. The Grade by Contextual Constraint and Grade by Case interactions were not significant, $F(2,234)=1.4$, $p > .05$ and $F(1,234)=.08$, $p > .05$ respectively.

DISCUSSION

To validate the measure proposed in the present paper, it was crucial to demonstrate that responses to word deletion items reflect the constraining influence of context. By controlling the influence of cues within the target sentence, the effect of increasing the information value of context on the uncertainty of the responses could be isolated. As contextual constraint is increased, the uncertainty of responses should decrease.

As the H values indicate, the uncertainty students had about the meaning of the target sentence was strongly affected by this experimental manipulation of context. A constraining context placed more restrictions upon the possible interpretations of the target sentence than the less constraining context, thus limiting the range of reasonable responses which could be given for a particular item. An example from the responses

of the sixth graders to one of the objectives items demonstrates this effect. For the highly constrained condition, 33 out of 40 responses to the item,

> His father fixed the broken wheel in the shed.
> John tried the _____ in the field.

involved mechanical things which have wheels and could be tested in the field, such as bicycles or tractors. When the word "wheel" in the context sentence was changed to the less constraining word "part," there were 30 responses concerning mechanical things or toys which might have broken parts and could be tested in a field (e.g., plow or kite). Only 18 of these responses, however, were objects which have wheels. In the low constraint condition where only the target sentence, "John tried the _____ in the field," was provided, only 24 responses involved things which had parts which might get broken and only eleven of these had wheels. The remaining 16 responses included food, objects to throw, and so on. Clearly, the semantic information in the context sentence was used to decrease uncertainty about meaning as reflected in the choice of words for the target sentence.

The lack of a difference and uncertainty between the Low and Medium levels of constraint was an interesting finding. A change in one word in the context sentence had a pronounced effect in the uncertainty of the responses whereas eliminating the context sentence had relatively little additional influence on uncertainty. Thus, the constraints implied by context must be rather explicit before young readers are able to use them. This underscores the importance of constructing highly constraining context for word deletion items.

As more fluent readers, sixth grade students should be better able than third graders to reduce uncertainty from context. Furthermore, if the word deletion items are primarily measuring the use of remote semantic context provided by the context sentence, this difference should dissipate as the constraint of content decreases. This relationship was clearly supported by the Grade by Contextual Constraint interaction. Since the only difference between the three forms of each item was the amount of semantic information available in context, the grade differences in uncertainty reduction observed at the High level of constraint appear to be due to the superior ability of sixth graders to utilize this contextual information. Had the sixth graders also constrained their answers more than third graders on the Low constraint items, the differences could have been interpreted in terms of sixth grade superiority in the use of syntactic or semantic cues in the target sentence.

It was also hypothesized that the number of words correct should decrease as the constraint of context increases. This hypothesis was supported by the large main effect due to Contextual Constraint. As constraint increases, the range of correct responses decreases and the reader must bring more information to bear in order to respond correctly. Differences in the scores on High constraint items could be attributed to individual differences in the ability to use this information and not other factors, since it was relatively easy for students to generate correct responses to the Low constraint items. The greater difficulty of the highly constrained items was due to the difficulty students encountered in using remote semantic context to comprehend when reading.

It was further hypothesized that the more fluent sixth graders would perform better than the third graders on the highly constrained items, but the advantage would be less when context was removed. This hypothesis was only partially supported by the data; there was a significant main effect for Grade but no interaction of Grade with Contextual Constraint. The failure to find an interaction may be due to a strong ceiling effect in the Low constraint condition which reduces the variability of the scores and increases the power of the test for Grade effects in that condition. However, it is possible that the difference between grades in that condition was also reduced by the ceiling effect. Thus, it would appear that the test scores reflect age differences in the ability to use contextual information but that the affects of reducing constraint of context could not be fully evaluated.

There are two potential alternative explanations of the age differences in both analyses. First, decoding problems may have been responsible for the lower performance of third graders. There may have been incomplete decoding (Guthrie & Tyler, Note 1) if the third graders could not identify some of the words or a lack of automatic control which would place demands on their limited information processing capacity (Perfetti & Hogaboam, 1975). Second, the poor performance might reflect a lack of relevant knowledge or experiences to make use of contextual information on particular items. These explanations seem unlikely, however, since only familiar words were used and the content of each item was limited to common, daily experiences that third graders might be expected to have.

The results of the present investigation suggest that the proposed contextually constrained word deletion procedure is valid, reliable and sensitive to individual differences. It is based on a definition of comprehension which has been shown to have considerable predictive and explanatory power, which clarifies the role of inference, and which emphasizes the importance of contextual dependence in measures of comprehension. Further evidence of construct validity is still needed to extend or replicate the present study, to investigate the relationship of this skill to other reading skills, and to determine the modifiability of the skill (e.g. an extension of Samuels, Dahl & Archwamety, 1974).

REFERENCE NOTES

1. Guthrie, J. T., & Tyler, S. J. Psycholinguistic processing in reading and listening among good and poor readers. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1976.

REFERENCES

Anderson, R. C. How to construct achievement tests to assess comprehension. Reading Research Quarterly, 1972, 42, 145-170.

Attneave, F. Application of information theory to psychology. New York: Holt, 1959.

Bormuth, J. R. On the theory of achievement test items. Chicago, Illinois: University of Chicago Press, 1970.

Bransford, J. D., & Johnson, M. K. Considerations of some problems of comprehension. In W. G. Chase (Ed.), Visual information processing. New York: Academic Press, 1973.

Carroll, J. B. Defining language comprehension: some speculations. In J. B. Carroll and R. O. Freedle (Eds.), Language comprehension and the acquisition of knowledge. Washington, D.C.: V. H. Winston & Sons, 1972.

Carroll, J. B., Davies, P. & Richman, B. Word frequency book. New York: American Heritage, 1971.

Chomsky, N. Aspects of the theory of syntax. Cambridge, Massachusetts: M.I.T. Press, 1965.

Davis, F. B. Research in comprehension in reading. Reading Research Quarterly, 1968, 3, 499-545.

Fillmore, C. J. The case for case. In E. Bach & R. T. Harms (Eds.), Universals in linguistic theory. New York: Holt, Rinehart & Winston, 1968.

Goodman, K. S. Reading: a psycholinguistic guessing game. Journal of the Reading Specialist, May, 1967.

Levin, H., & Williams, J. P. (Eds.) Basic studies on reading. New York: Basic Books, 1970.

MacGinitie, W. H. Contextual constraint in English prose paragraphs. Journal of Psychology, 1961, 51, 121-130.

Olson, D. R. Language and thought: aspects of a cognitive theory of semantics. Psychological Review, 1970, 77, 257-273.

Perfetti, C. A. Psychosemantics: some cognitive aspects of structured meaning. Psychological Bulletin, 1972, 78, 241-259.

Perfetti, C. A., & Hogaboam, T. Relationship between single word decoding and reading comprehension skill. Journal of Educational Psychology, 1975, 67, 461-469.

Pyrczak, F. Passage dependency of items designed to measure the ability to identify the main ideas of paragraphs: implications for validity. Educational and Psychological Measurement, 1974, 34, 343-348.

Samuels, S. J., Dahl, P., & Archwamety, T. Effect of hypothesis/test training on reading skill. Journal of Educational Psychology, 1974, 66, 835-844.

Smith, F. Understanding reading. New York: Holt, Rinehart and Winston, 1971.

Spearitt, D. Identification of subskills of reading comprehension by maximum likelihood factor analysis. Reading Research Quarterly, 1972, 8, 92-111.

Taylor, W. L. "Cloze" readability scores as indices of individual differences in comprehension and aptitude. Journal of Applied Psychology, 1957, 41, 19-26.

Thorndike, R. L. Reading as reasoning. Reading Research Quarterly, 1973-4, 9, 135-147.

Vaughan, G. M., & Corballis, M. C. Beyond tests of significance: estimating strength of effects in selected ANOVA designs. Psychological Bulletin, 1969, 72, 204-213.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.